The CDC's Controlled Health Thesaurus and the Unified Medical Language System

Edward Bunker, MPH
Olivier Bodenreider, MD, PhD

July 29, 2005

Abstract

The Controlled Health Thesaurus (CHT) is a controlled vocabulary that was developed by the Centers for Disease Control and Prevention (CDC) to support the automated indexing and retrieval of public health content on the CDC web site. The CHT evolved from four vocabularies contained in the Unified Medical Language System (UMLS) of the National Library of Medicine (NLM). As the CHT has evolved, terms have been added to both the CHT *and* the UMLS. We undertook a study to identify those terms in the CHT which could be mapped to the UMLS but which currently do *not* have an assigned UMLS Concept Unique Identifier (CUI). We employed automated methods to generate, display and semantically validate proposed UMLS CUI mappings. Our efforts identified 408 terms in the CHT that have semantically valid UMLS CUI mappings, and we propose that these mappings could be considered for inclusion in the CHT.

Introduction and Background

The Unified Medical Language System® (UMLS®) of the National Library of Medicine (NLM) was developed to provide access to and linkages between the many controlled vocabularies that exist in the biomedical domains. In a 1996 study of the coverage of biomedical concepts in the UMLS, Humphreys stated, "…the use of controlled vocabulary in health care and public health systems is likely to increase the quality, effectiveness, and efficiency of health care and to facilitate clinical research, public health surveillance, and health services research." (Humphreys, et al, 1997). Among the many task areas covered in this study was public health, but only 8% of the terms submitted for consideration in this study concerned "public health surveillance or intervention." Terms could be categorized under multiple areas, and a majority of terms (74%) were considered to fall under the task area of "Direct patient care." The authors reported that of the 3,239 public health terms submitted, 50% were found to semantically match with concepts found in the UMLS, and 49% were found to match with related concepts. (Humphreys, et al, 1997). A review of the literature on biomedical terminology in Medline suggests that the Humphreys study is the only one that has ever addressed coverage of public health terms in the UMLS.

Public health is distinct from medicine. While there are some commonalities in approach and practice, public health practitioners communicate to their peers and to the public using a different set of concepts, terms and assumptions. To index and improve access to public health content on

their web site (www.cdc.gov), the CDC has developed a specialty vocabulary that attempts to express the unique perspectives and needs of public health community. (Bell, 2005a) This vocabulary, referred to as the Controlled Health Thesaurus (CHT), is part of a suite of controlled vocabularies maintained and supported by the CDC. These vocabularies are part of CDC's overall effort to promote and promulgate data standards under the umbrella of the Public Health Information Network (PHIN). (Bell, 2005b)

The CHT has its genesis in four source vocabularies of the National Library of Medicine's (NLM) Unified Medical Language System (UMLS). These four source vocabularies are: NLM's Medical Subject Headings (MeSH), Computerized Retrieval of Information on Scientific Projects (CRISP), Alcohol and Other Drug Thesaurus (AOD), and the Library of Congress Subject Headings (LCH). The developers of the CHT are planning to include SNOMED-CT terms in their vocabulary as it was released for use in the US in 2003. (Bell, 2005a) One of the reasons these source vocabularies were chosen to seed the CHT was that under the current UMLS licensing agreement these sources can be distributed and shared with minimal constraints. (Bell, 2005a) However, a more important reason they were chosen is that they form a solid basis upon which a biomedical vocabulary can be developed. "…much of the vocabulary needed for health data is already available and…a combination of existing systems may provide the best foundation for building a comprehensive clinical vocabulary." (Humphreys, et al, 1997)

The technical work for the CHT has been contracted to Kevric, Inc. Apelon tools are being used for development.

Research Questions

We were motivated to pursue several areas of inquiry. First, upon a preliminary examination of the CHT we noted that several fairly common biomedical terms did not have an assigned UMLS Concept Unique Identifier (CUI). Why did these terms not have an assigned CUI? Were there actual conceptual differences between the terms as they hierarchically reside in the CHT and the UMLS? Was the level of specificity available in the UMLS for certain concepts insufficient for public health? Were there errors of CUI omission during the CHT development or editing process? Finally, would licensing or proprietary considerations account for the lack of UMLS CUI mapping?

As a consequence of attempting to answer these questions, a second area of inquiry emerged. Questions in this area included: As the UMLS and UMLS-*based* vocabularies evolve (i.e., add, delete and change terms, relationships, etc.) what sorts of processes and tools might be useful for assessing coverage of "new" terms and relationships? How can coverage information be presented to vocabulary developers and domain experts to facilitate vocabulary curation.

Materials and Methods

Procurement of the CHT Vocabulary

In July of 2005 we downloaded an Excel version of the CHT from the CDC web site. (http://www.cdc.gov/phin/vocabulary/). The CHT came in a "flat" format. Some of the fields included: Description, Base Code System, Parent Code Name, Parent Code System, status,

Apelon Concept Name, Semantic Type, Scope Note, UMLS Concept Unique Identifier (CUI), Associated MeSH Concept, Associated CRISP Concept, Associated AOD Concept, Associated ICD-9-CM Concept, Associated HL7 Concept, Parent Of String, Child Of String, and Synonym String. For the purposes of our study we extracted the following fields: Description (or Term), UMLS CUI, Parent-Of String, and Semantic Type.

Identification and Selection of Terms

Of the 42,639 terms contained in the CHT, 30,639 had one or more assigned UMLS CUI's. For instance, the CHT term "Adrenalin" had a UMLS CUI designation of "C0014563." After removing the terms with a UMLS CUI, we were left with a set of 12,000 CHT terms. Of the 12,000 terms that did *not* have a UMLS CUI, 7,257 had a semantic type of "Geographic Area" or "Partner Organization." While these semantic types are important for the purposes of the CHT, we decided to not include them in our subsequent analyses because they were of more general interest and not specific to public health practice per se.

For our analysis, we had a set of 4,743 CHT terms. Examples of these terms included: "Used Car Salesman," "Alpha Radiation," and "Italian."

Mapping of CHT Terms to UMLS Concepts

We used a "normalized string index" approach to map between CHT terms and UMLS concepts. Our original approach – which we eventually abandoned – entailed an "exact match" method followed by a normalized string index approach; however, a casual examination of initial results revealed that too few candidates were being adequately mapped. For instance, using our original approach (exact match followed by normalized string index) the CHT term "Italian" only mapped to the UMLS concept for "Italian Language." Knowing that this was an invalid mapping and that the UMLS indeed contained a concept of "Italians" (the population), we employed only the normalized string index approach. Under the normalized string index approach the CHT term "Italian" mapped to both "Italian Language" and "Italians". While we know that a normalized string index approach would require more effort to disambiguate, we were willing to pay this price for the benefit of capturing more potential candidate concepts.

The automated mapping procedures were performed with customized tools developed by one of the authors (OB) using Perl.

Semantic Validation

We sought to semantically validate our mappings through a two step process of "*establishing the lineage of a CHT term*" and "*identifying common ancestor*s." In cases where multiple mappings were produced, we also performed "*disambiguation*."

Establishing the lineage of a CHT term: To understand the contextual meaning of a particular term within the CHT hierarchy, we produced an ancestor mapping for each term of interest. The mappings were derived from the "Parent" relationships defined in the CHT. For instance, "Alpha radiation" has "Ionizing radiation" defined as a parent; "Ionizing radiation" has "Physical phenomenon" defined as a parent, etc. The complete CHT mapping for "Alpha radiation" is displayed below:

CDC Controlled Health Thesaurus
>Processes and phenomena
>Physical phenomenon
>Ionizing radiation
>Alpha radiation: *Proposed UMLS Concept "Alpha Particles" (C0002217)*

This lineage (or hierarchy) was proposed for manual review, and an editor might reasonably be able to validate, for instance, that "Alpha radiation" in CHT relates to "Alpha particles" in UMLS. However, a manual examination of the hierarchy of a term in order to gain knowledge about its context did not seem practical for even a cursory review of a very large number of concepts. We therefore proposed a second complementary approach to automate the process of semantic validation.

Identifying common ancestors: To provide further semantic validation for a proposed mapping, we identified which ancestors of a term might have a corresponding UMLS CUI assigned. In the below example, "Ionizing radiation" (an ancestor to the term of interest, "Alpha radiation") has a UMLS CUI of C0034538.

CDC Controlled Health Thesaurus
>Processes and phenomena
>Physical phenomenon
>Ionizing radiation *(UMLS CUI C0034538)*
>Alpha radiation: *Proposed UMLS Concept "Alpha Particles" (C0002217)*

For further semantic validation, the question can be asked: Is the proposed UMLS Concept (in this case "Alpha radiation"), a descendant of any of the CHT ancestor *concepts*? In this example, the answer is YES: Alpha particles is a descendant of "Ionizing radiation." Given that the mapping between "Alpha radiation" and "Alpha Particles" was unique, and given that an ancestor for "Alpha Particles" could be found in the CHT lineage of "Alpha radiation", we declared this mapping to be "semantically valid." To identify ancestors of proposed UMLS concepts, we used a pre-computed list of ALL ancestors for ALL UMLS concepts, created from a slightly modified version of the UMLS Metathesaurus. In this version, hierarchical circular relations have been removed. This list was computed by one of the authors (OB).

Disambiguation: For mappings that produced multiple matches (e.g., "Italian" mapping to "Italians" and "Italian Language"), semantic validation was used essentially for disambiguation purposes. Essentially this entailed choosing the mapping that identified the greatest number of valid ancestor matches.

Population Group By Race
>White (C0043157)
>European
>Italian: *Proposed UMLS Mappings:*
#1) Italian Language, C0022275 (0 Ancestors)
#2) Italians, C0337810 (1 Ancestor, best candidate)

In the case of the CHT term “Italian”, the mapping to the UMLS concept “Italians” (CUI C0337810) was chosen because the concept “White” (CUI C0043157) was found to be its ancestor. “Italian Language” (CUI C0022275) had no identified UMLS ancestors in the CHT lineage.

In cases where there were equal numbers of valid ancestor matches, we arbitrarily chose the first candidate in the series to break the tie.

Results

Of the 4,743 CHT terms that we attempted to match to UMLS concepts, 3,900 (82.2%) returned no UMLS concept matches, while 843 (17.8%) returned 1 or more UMLS concept match.

Of these 843 matches, 744 matches were “unique” (i.e., the CHT term only mapped to 1 UMLS concept.) The other matches, totaling 99, produced multiple UMLS concept matches. A summary of results is detailed in Table 1.

Table 1. Results of UMLS concept mapping and Semantic Validation.

843 CHT Terms Returned UMLS Matches			
744 Unique Matches		99 Multiple Matches	
354 ***Semantically Valid***	390 Not Semantically Valid	***54*** ***Semantically Valid***	60 Not Semantically Valid

A total of 408 matches were considered to be semantically valid, while 435 were determined to not be semantically valid. Examples from our automated process are presented below. Please note that semantic validity is not determined by simple matching of strings. For example, the CHT term “Chloroflexi” does not validly match with the UMLS Concept of “Chloroflexi” because no common ancestors were identified. A human editor might reasonably override this determination.

Examples of Unique Matches, Semantically Valid

CHT term “Body fat distribution” and UMLS Concept “*Distribution of body fat*”

CHT term “Bone cancer” and UMLS Concept “*Malignant Bone Neoplasm*”

Examples of Unique Matches, *Not* Semantically Valid

CHT term “Bone set” and UMLS Concept “*Skeletal bone*”

CHT term “Breda virus” and UMLS Concept “*Bovine torovirus*”

Examples of Multiple Matches, Semantically Valid

CHT term “Bone mineralization” and UMLS Concept “*Physiologic calcification”* (Valid)
CHT term “Bone mineralization” and UMLS Concept “*Ossification, Physiologic*” (*Not* Valid)

CHT term “Chinese” and UMLS Concept “*Chinese People*” (Valid)
CHT term “Chinese” and UMLS Concept “*Chinese Language*” (*Not* Valid)

Examples of Multiple Matches, Not Semantically Valid

CHT term “Chloroflexi” and UMLS Concept “*Chloroflexi*” (*Not* Valid)
CHT term “Chloroflexi” and UMLS Concept “*Chloroflexus*” (*Not* Valid)

CHT term “Continent” and UMLS Concept “*Continent*” (*Not* Valid)
CHT term “Continent” and UMLS Concept “*Geographic continent*” (*Not* Valid)

Source Vocabularies of “Recovered” Concepts

Using resources of the UMLS Knowledge Source Server, we determined the source vocabularies for the 408 semantically valid UMLS concept matches. It should be noted that UMLS concepts can have more than 1 source vocabulary. The top five sources for English names of these “recovered” concepts are detailed in Table 2.

Table 2. Top Five Source for English Names of “Recovered” Concepts

Source Vocabulary	% of Recovered Terms
SNOMED-CT	75%
SNOMED International	56%
Read Codes	48%
MeSH	31%
CRISP	22%

Limitations

We note several limitations to our study. First, we performed our normalized string matching using data from the 2005 version of the UMLS, while our ancestry identification was performed using 2004 data. This versioning variation probably lowered the percentage of semantically valid mappings.

Second, we did not manually review each mapping to assess the "face" validly of this automated process. In an actual production environment, however, the type of output we produced could be used in a manual curation process.

Discussion

We have shown that it is technically feasible to facilitate a process of "resynchronization" of UMLS-based biomedical vocabularies. We believe that the processes we have outlined here could help re-establish conceptual linkages between vocabularies that have evolved apart over time.

In the future, we would like to study the 10% (or 27%) of CHT terms that did not validly map to any UMLS concepts. We would propose to perform a frequency analysis of ancestor terms to discover what concept areas might not be well represented in the UMLS. In this way the UMLS could be strengthened by the public health perspective that is contained within the CHT. We would also propose that our method could be used to validate all *existing* CUI assignments in the CHT and other vocabularies, and in this way our method could serve as a further check of the structure and integrity of these vocabularies. We would also like to assess whether our display of hierarchical context would be a useful tool for developers, editors and curators of controlled vocabularies in the biomedical and other sciences.

Finally, normalized string matching was shown to be a preferred method for our purposes as it helped to identify a greater number of candidate mappings to the UMLS.

References

Bell M. [2005a] CDC's Controlled Health Thesaurus: Common language to identify and describe information. CDC Web Site (http://www.cdc.gov/phin/vocabulary/Controlled_Health_Thesaurus_Brochure_v2.1.pdf) last accessed July 28, 2005;1-2.

Bell M. [2005b] Personal Communication. July 5, 2005.

Humphreys B, McCray A, and Cheh M. Evaluating the Coverage of Controlled Health Data Terminologies: Report on the Results of the NLM/AHCPR Large Scale Vocabulary Test. J Am Med Inform Assoc. 1997;4:484-500.